WAVEFORMS AND SPECTROGRAMS: ENHANCING ACOUSTIC SCENE CLASSIFICATION USING MULTIMODAL FEATURE FUSION

Dennis Fedorishin[1], *Nishant Sankaran*[1], *Deen Dayal Mohan*[1], *Justas Birgiolas*[2,3]
Philip Schneider[2], *Srirangaraj Setlur*[1], *Venu Govindaraju*[1]

[1] University at Buffalo, Center for Unified Biometrics and Sensors, USA,
{dcfedori, n6, dmohan, setlur, govind}@buffalo.edu
[2]ACV Auctions, USA, {jbirgiolas, pschneider}@acvauctions.com
[3]Ronin Institute, USA.

ABSTRACT

Acoustic scene classification (ASC) has seen tremendous progress from the combined use of convolutional neural networks (CNNs) and signal processing strategies. In this paper, we investigate the use of two common feature representations within the audio understanding domain, the raw waveform and Mel-spectrogram, and measure their degree of complementarity when using both representations for feature fusion. We introduce a new model paradigm for acoustic scene classification by fusing features learned from Mel-spectrograms and the raw waveform from separate feature extraction branches. Our experimental results show that our proposed fusion model significantly outperforms the baseline audio-only subnetwork on the DCASE 2021 Challenge Task 1B (increase of 5.7% in accuracy and a 12.7% reduction in loss). We further show that the learned features of raw waveforms and Mel-spectrograms are indeed complementary to each other and that there is a consistent improvement in classification performance over models trained on Mel-spectrograms or waveforms alone.

Index Terms— Audio classification, Acoustic scene classification, Feature fusion, Multi-modal features.

1. INTRODUCTION

Mel-spectrograms are the de-facto audio feature representation and have been widely used throughout the history of audio understanding [1]. Mel-spectrograms are created by calculating the short-time fourier transform (STFT) of an audio signal, then passing the STFT frequency responses through band-pass filters spaced on the Mel(logarithmic)-scale and often further passed through a logarithmic compression to replicate the human's non-linear perception of signal pitch and loudness, respectively.

With the advent of deep neural networks, many methods have been introduced that perform audio understanding tasks such as ASC, audio tagging, and sound event detection by using Mel-spectrogram representations of audio as the input to a convolutional neural network [2, 3]. Researchers have also explored other feature representations such as the gammatone and Constant-Q (CQT) spectrogram, and Mel Frequency Cepstrum Coefficients (MFCC) [4, 5]. [6] and found that fusing these representations allows for a network to learn complementary features, creating a stronger model for ASC.

In parallel, other works have utilized the raw waveform directly as input into neural networks, bypassing the need for hand crafted features [7, 8]. Waveform-based networks are trained end-to-end, while networks that utilize spectrograms need to create these hand crafted features that may often be sub-optimal for the given task. Regardless, many state of the art methods in ASC, speaker recognition, sound event detection, and other tasks still utilize spectrogram representations [9, 10]. Further, [11] introduced a fully learnable variation of spectrogram representations, where they are trained end-to-end to automatically find an optimized representation.

As a result, there is still no clear consensus as to the best feature representation that can perform strongly across various audio understanding tasks. Researchers are now looking at hybrid methods that use both waveform and spectrogram representations in a fusion setting. [8, 12] perform early feature map fusion of waveform and spectrogram features that are passed through convolutional layers for audio tagging and environmental sound classification. [13, 14] propose a decision-level ensembling of multiple models that utilize raw waveforms and Mel-spectrograms for environmental sound classification and ASC. Although these works have shown classification performance improvements using waveforms and spectrograms, they do not deeply explore the degree of complementarity and effects of fusing these features together.

In this paper, we investigate waveform and Mel-spectrogram feature fusion and propose a new ASC model that learns complementary features from both modalities using a more effective fusion method. We evaluate our proposed model using the DCASE 2021 Challenge Task 1B dataset to prove the effectiveness and complementarity of waveform and Mel-spectrogram feature fusion. Our work is reproducible and the code is publicly available.[1]

2. PROPOSED METHOD

To investigate and understand the complementarity between learning features from Mel-spectrograms and raw waveforms, we designed a fusion model based on two CNN feature extractors, and a unified classification layer. Figure 1 illustrates the design of our model. The spectrogram branch, F_s, is comprised of repeating 2D CNN blocks followed by a max pooling operation. The CNN blocks contain a convolution layer using a kernel size of 3×3, followed by a batch normalization and a Leaky ReLU nonlinear activation.

The waveform branch, F_w, is of a similar structure, however the two-dimensional max pooling and convolutional layers are replaced with one-dimensional kernels of size 8 and 7, respectively. In addition, the first convolutional layer in the waveform branch are parameterized to Sinc functions, as described in [15].

[1] https://github.com/denfed/wave-spec-fusion

Table 1: Detailed overview of proposed model design.

Spectrogram Branch F_s	
Input shape	[128,188]
Filter responses	32, 64, 128, 256
Global average pooling output l_s	$\langle 1024 \rangle$
Waveform Branch F_w	
Input shape	[48000]
Filter responses	32, 64, 128, 256
Global average pooling output l_w	$\langle 1024 \rangle$
Classification Layers F_c	
Input shape $(l_w + l_s)$	$\langle 1024 \rangle$
Dense layer outputs	$512, 256, 10$
Output classes	10

As both branches are working with different-sized input data, the feature responses may vary in size. We utilize global average pooling layers to condense both waveform and spectrogram features into a vector of 1024 units, denoted by l_w and l_s, respectively.

Feature fusion of both feature extraction branches is accomplished at the latent representation level, where features for both the waveform and spectrogram branch are extracted independently and then fused together into a unified representation. We hypothesize that this style of late fusion allows each feature extractor to extract more complementary features compared to related methods that fuse feature maps in early layers. In addition, late fusion may exploit more feature inter-dependencies of each modality compared to score fusion and ensembling methods that average independent model outputs.

Fusion is accomplished using element-wise summation such that the final latent representation $(l_w + l_s)$ has the same shape as its constituents. The classification layers, F_c, take $(l_w + l_s)$ as input and perform the final classification using repeating dense blocks, as shown in Figure 1. We use dropout layers with $p = 0.3$, followed by linear layers, a Leaky ReLU activation, and batch normalization. The classification $\hat{c}$ of the set of classes $c \in \mathbb{C}$ of an audio sample with its waveform x_w and Mel-spectrogram x_s can be described as:

$$\hat{c}_{(x_w, x_s)} = \underset{c \in \mathbb{C}}{\operatorname{argmax}} F_c(F_w(x_w) + F_s(x_s)) \quad (1)$$

Table 1 describes in detail the configuration of both feature extraction branches and the final classification layers. For our experiments, we utilize three variations of the described model to investigate modality complementarity:

Spectrogram sub-network: The spectrogram branch in Figure 1 is used independently with the classification layers, without the waveform branch. In this model, training is conducted only using Mel-spectrograms, omitting $F_w(x_w)$ from (1).

Waveform sub-network: The waveform branch in Figure 1 is used independently with the classification layers, without the spectrogram branch. In this model, training is conducted only using raw waveforms, omitting $F_s(x_s)$ from (1).

Fusion model: Both the spectrogram and waveform branch are trained end-to-end with their respective inputs. The latent representations of each branch are fused together for classification.

3. TRAINING CONFIGURATION

All models are trained using the SGD optimizer paired with the one-cycle learning rate scheduler described in [16]. The one-cycle learning rate scheduler anneals an initial learning rate to a maximum value, then anneals it back to the initial learning rate, over the entire training procedure. [17] showed that this procedure leads to faster training times. In addition, using large learning rates for a portion of the training procedure acts as a form of regularization. We experimented with various learning rates and found that a max learning rate of 0.008 for the one-cycle scheduler works best for the proposed models. Training batch size is set to 128 and the models are trained for 50 epochs. The models were trained on an RTX 6000 GPU with the most complex model taking 1.5 hours to fully train. We train our models and conduct our experiments using the provided training and validation split of the Task 1B challenge dataset.

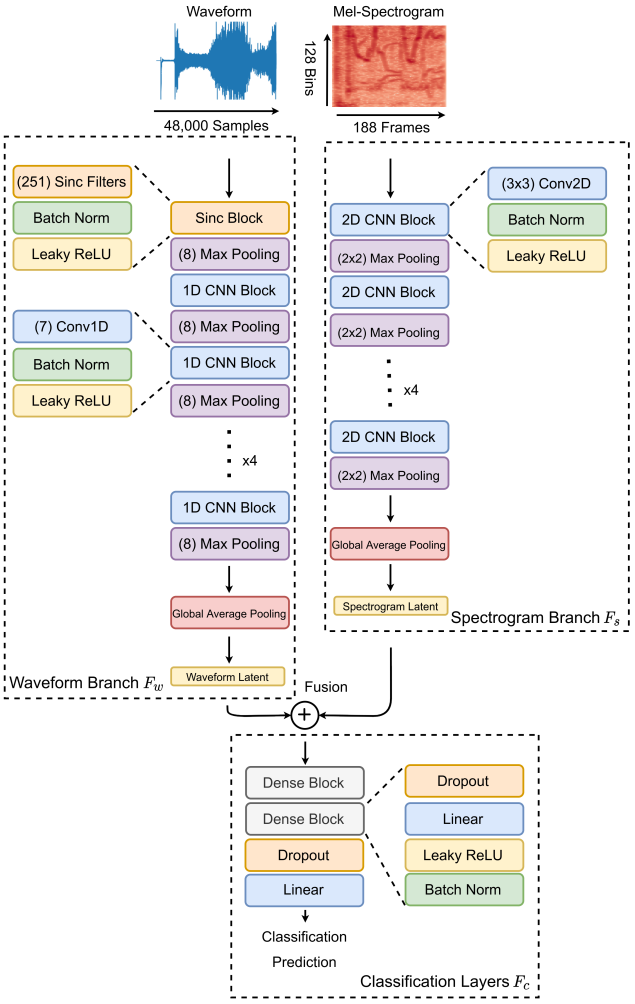

Figure 1: Illustration of the proposed fusion model.

3.1. Dataset: DCASE 2021 Challenge Task 1B, Audio-Visual Scene Classification

Task 1B is based on the TAU Audio-Visual Urban Scenes 2021 dataset, a dataset containing synchronized audio and video recordings from 12 European cities in 10 different scenes. Audio is recorded using a Soundman OKM II Klassik/studio A3 microphone paired with a Zoom F8 audio recorder, sampled at $48kHz$ at a 24-bit resolution. Video is recorded using a GoPro Hero5 Session. The

Table 2: Kernel parameterization performance.

Parameterization	Accuracy %	Log Loss
Unparameterized (normal)	61.87	1.047
Sinc parameterization [15]	**64.79**	**1.045**

Table 3: Model performance compared to challenge baseline.

Model	Accuracy %	Log Loss	# Params
Audio baseline [20]	65.1	1.048	-
Waveform sub-network	64.79	1.045	1.0M
Spectrogram sub-network	66.46	1.072	1.1M
Fusion Model	**70.78**	**0.915**	1.4M

dataset contains 12,292 10-second samples of each modality spread across the 10 scenes. The provided train/validation split consists of 8,646 samples in the training set and 3,645 samples in the validation set [18]. In this paper, we focus on the audio portion of the dataset only, excluding using videos for scene classification.

3.2. Data Preprocessing

We input the raw waveform and its generated Mel-spectrogram into their respective feature extractors. According to the Task 1B rules, we split the development dataset samples into 1 second audio files to perform classification at the 1 second level. This brings the training dataset to 86,460 samples and the validation dataset to 36,450 samples. Audio files are sampled at $48kHz$ and therefore have a sample length of $[48000]$. In addition, the audio waveforms are scaled to the range $[0, 1]$. Mel-spectrograms are generated using 128 frequency bins, a hop length of 256 samples, and a Hann window size of 2048 samples, creating a final size of $[128 \times 188]$. The Mel-spectrograms are also passed through a logarithmic compression and then normalized at an instance level using Z-Score normalization such that each sample has a mean of 0 and unit standard deviation.

3.3. Data Augmentation

For data augmentations, we utilize Mixup [19] and time shifting for all experiments. During training, Mixup has a 50% probability of being used for each batch and time shifting is applied to every batch. For Mixup, we select $\alpha = 0.2$ and apply it to both the sub-networks and fusion model. For the fusion model, Mixup is performed evenly across the waveform and spectrogram such that two audio samples' waveform and spectrogram are mixed together at the same mixing ratio. Time shifting shifts the waveform and spectrogram along the time axis, where overrun samples are shifted to the opposite end of the input. We time shift randomly from 0% to 50% of the time axis size for both the waveform and spectrogram. For the fusion model, time shifting is applied independently to the modalities, such that both the waveform and spectrogram may be shifted by varying degrees. We experimented with various configurations and found that this configuration achieves the highest classification performance, however further research should be conducted on the effects of consistent and inconsistent data augmentations between each modality.

4. EXPERIMENTAL RESULTS

4.1. Waveform Kernel Parameterizations

SincNet [15] introduced the use of parameterized Sinc filters for speech recognition, where the kernels of the first convolutional layer of a model utilizing waveforms are replaced with kernels parameterized to the Sinc function. Works such as [14] have applied these filters for ASC and found Sinc filters to improve ASC performance.

Table 4: Fusion method comparisons.

Model	Accuracy %	Log Loss	# Params
Wavegram-Logmel-CNN	68.35	1.063	80.2M
Decision fusion	68.65	0.955	2.0M
Decision ensemble	70.47	**0.845**	2.0M
Proposed late fusion	**70.78**	0.915	1.4M

Table 2 shows model performance when replacing the first convolutional layer of the waveform branch with parameterized Sinc kernels instead of an unparameterized, fully learnable kernel. As shown, using sinc kernels outperforms unparameterized kernels. We hypothesize that the initialization of the sinc filters to mirror the distribution of the Mel-scale, as described in [15], are a more optimal initialization compared to conventional kernel initializations. In addition, the Sinc kernels are less prone to overfitting as they are constrained to the Sinc function [11]. Using Sinc filters also allows us to reduce model complexity, as two frequency cutoff values are learned per kernel instead of N parameters of a size N kernel.

4.2. Waveform and Spectrogram Feature Fusion

Table 3 shows the classification performance of the provided Task 1B baseline model compared to the three different model variations proposed. The waveform sub-network is not able to outperform the baseline while the spectrogram sub-network performs slightly better than the baseline in accuracy. The fusion model outperforms both the baseline and models trained on single modalities with a 5.7% improvement in accuracy and a reduction of .13 in loss over the baseline. Furthermore, we see that the fusion model outperforms the spectrogram sub-network by 4.3% in accuracy and a .16 reduction in loss. This improvement shows that there are features being learned within the raw waveform that are complementary to features being learned from the Mel-spectrogram, resulting in a more discriminative classification model.

4.3. Feature Fusion Design

We perform a comparison with other fusion paradigms to better understand its significance in fusing waveform and spectrogram features. We compare our method against the Wavegram-Logmel-CNN, a popular acoustic classification model introduced by [12] that performs early feature map fusion on the waveform and spectrogram. We train the Wavegram-Logmel-CNN using the training configuration described in [12]. The same data preprocessing is used as described in section 3.2, however the spectrogram hop length is changed to 320 samples to fit the structure of the model. In addition, we compare the proposed late fusion design against a decision fusion and ensembling method. Instead of latent vector fusion, we fuse the independent sub-network's predictions at the decision level by averaging predictions together. Comparing to (1), the decision fusion classification resembles:

$$\hat{c}_{(x_w, x_s)} = \underset{c \in \mathbb{C}}{\operatorname{argmax}} \frac{1}{2}(F_{c_w}(F_w(x_w)) + F_{c_s}(F_s(x_s))) \qquad (2)$$

where F_{c_w} and F_{c_s} depict the waveform and spectrogram subnetwork's classification layers, respectively. We train this decision fusion model using the same configuration as the proposed late fusion model. In addition, we compare the fusion methods against an ensemble of the independently trained sub-networks, using decision averaging described in (2).

As shown in Table 4, the proposed late fusion model outperforms the Wavegram-Logmel-CNN with significantly fewer parameters, in addition outperforming the decision fusion and ensemble

Table 5: Comparison of feature fusion methods.

Fusion Method	Accuracy %	Log Loss	# Params
Element-wise sum	70.78	**0.915**	1.4M
Concatenation	**70.85**	0.924	1.9M
MFB [21]	70.13	0.943	7.6M

Table 6: Feature branch removal ablation study.

Model	Accuracy %	Log Loss
Spectrogram sub-network	66.46	1.072
Fusion spectrogram branch only	51.33	1.720
Waveform sub-network	64.79	1.045
Fusion waveform branch only	31.51	2.500

model. We see that while the ensemble of both sub-networks performs well, the late fusion model is able to extract feature interdependencies of the waveform and spectrogram that still outperform the ensemble model in terms of accuracy. The late fusion model also has fewer parameters and is trained end-to-end.

4.4. Latent Representation Fusion Methods

Most approaches to feature fusion utilize linear methods such as element-wise summation and concatenation of vectors and feature maps. A more advanced operation, Multimodal Factorized Bilinear Pooling [21], has been used within visual question answering and captures more expressive features than linear methods while being less computationally expensive than conventional bilinear pooling.

We experiment using these fusion methods to see whether we can fuse features in a more expressive fashion. Table 1 and Figure 1 depict the design for element-wise summation fusion. For concatentation, latent vectors l_w and l_s are combined to a final size of 2048 units. This new vector is passed into the classification layers, with the dense layers outputting $1024, 512, 10$ units, respectively. For MFB fusion, we set $k = 3$ and $o = 1024$, as described in [21]. The MFB fusion model has the same design as Figure 1, but the element-wise summation operation is replaced with MFB.

Table 5 shows the performance of our fusion model when utilizing element-wise sum, concatenation, and MFB. All methods perform similarly, however element-wise summation produces the lowest validation loss model. Fusion by concatenating latent vectors results in the highest accuracy model. We select element-wise summation fusion as it produced the lowest loss in addition to it being the least computationally expensive operation.

5. ABLATION STUDIES

Although we see a classification performance improvement when fusing waveform and spectrogram features, we must validate that the improvement is from complementary features extracted from both modalities. It may be the case that the feature extraction branches are underparameterized, and when adding more parameters the model performs better solely due to the increase in parameterization and not the second modality. To test this hypothesis, we expand both sub-networks such that their number of parameters exceed the fusion model by doubling each of the CNN block filter responses, classification layer responses, and increasing latent vectors to 2048 units. As shown in Table 7, both sub-networks were unable to surpass the performance of the fusion model, showing that the added performance in the fusion model is from the added modality.

To further understand the differences of each sub-network's performance, we compare each sub-network to their equivalent sub-network trained in the fusion setting. Examining the performance drop when removing feature extraction branches F_w and F_s in the fusion model may give clues into how the branches train alone versus in the fusion setting. The trained waveform and spectrogram sub-networks depicted in Table 3 are compared to the fusion model's respective sub-network. As shown in Table 6, the sub-networks that are trained in the fusion setting have a substantial performance loss when removing the opposite sub-network, far below the performance of the respective sub-network that is trained independently. We infer that when trained end-to-end, each of the sub-networks in the fusion model learn to focus on disparate features that when fused together, improve classification performance.

Table 7: Parameterization ablation study.

Model	Accuracy %	Log Loss	# Params
Fusion model	**70.78**	**0.915**	1.4M
Large spectrogram sub-network	66.48	1.043	4.2M
Large waveform sub-network	63.44	1.041	3.9M

Table 8: Class-Wise losses of the fusion model.

Class-Wise loss	Fusion	Fusion spec. branch only	Fusion wave. branch only
Airport	**0.901**	1.226	3.441
Shopping Mall	**0.944**	0.995	1.612
Metro Station	**1.053**	2.030	1.827
Street Pedestrian	1.104	**1.069**	2.638
Public Square	1.321	1.384	**0.663**
Street Traffic	**0.424**	0.843	3.038
Tram	**0.899**	2.106	4.182
Bus	0.747	4.905	**0.723**
Metro	**1.145**	2.825	4.324
Park	0.535	**0.443**	2.913

We also investigate class-wise loss when removing each branch of the fusion model, as shown in Table 8. Most classes have the lowest loss in the fusion model, however when removing the waveform branch, the spectrogram branch has a lower loss for the Street Pedestrian and Park class. When removing the spectrogram branch, the waveform branch has a lower loss for the Public Square and Bus class. We infer that while the fusion model can generally capture complementary features from each modality, the fusion operation is not able to exploit the full degree of complementarity of each branch's features. A fusion method that can fully exploit the modality complementarity would further improve ASC performance.

6. CONCLUSION

In this paper, we investigate feature fusion of two common audio representations, the raw waveform and Mel-spectrogram, and show that there are complementary features being learned that improve ASC performance. Further, we explore various fusion methods and experimentally validate that the proposed late fusion model is able to outperform other feature fusion designs. Our proposed fusion model utilizes these features to significantly outperform the DCASE 2021 Challenge Task 1B audio baseline and achieve 2$^{\text{nd}}$ place against the audio-only submissions. In future work, we will investigate more fusion methods to better exploit waveform and spectrogram feature complementarity and explore the effects of using independent data augmentations on the separate modalities.

7. ACKNOWLEDGMENT

This work was supported by the Center for Identification Technology Research (CITeR) and the National Science Foundation (NSF) under grant #1822190

8. REFERENCES

[1] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. Sainath, "Deep learning for audio signal processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 206–219, 2019.

[2] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, *et al.*, "CNN architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 131–135.

[3] N. Turpault, R. Serizel, A. Parag Shah, and J. Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *2019 Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, New York City, United States, October 2019.

[4] L. Pham, H. Phan, T. Nguyen, R. Palaniappan, A. Mertins, and I. McLoughlin, "Robust acoustic scene classification using a multi-spectrogram encoder-decoder framework," *Digital Signal Processing*, vol. 110, p. 102943, 2021.

[5] Y. Su, K. Zhang, J. Wang, D. Zhou, and K. Madani, "Performance analysis of multiple aggregated acoustic features for environment sound classification," *Applied Acoustics*, vol. 158, p. 107050, 2020.

[6] H. Wang, Y. Zou, and D. Chong, "Acoustic scene classification with spectrogram processing strategies," in *2020 Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, Tokyo, Japan, November 2020, pp. 210–214.

[7] T. Kim, J. Lee, and J. Nam, "Sample-level cnn architectures for music auto-tagging using raw waveforms," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 366–370.

[8] B. Zhu, C. Wang, F. Liu, J. Lei, Z. Huang, Y. Peng, and F. Li, "Learning environmental sounds with multi-scale convolutional neural network," in *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1–8.

[9] K. Miyazaki, T. Komatsu, T. Hayashi, S. Watanabe, T. Toda, and K. Takeda, "Conformer-based sound event detection with semi-supervised learning and data augmentation," in *2020 Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, Tokyo, Japan, November 2020, pp. 100–104.

[10] L. Sarı, K. Singh, J. Zhou, L. Torresani, N. Singhal, and Y. Saraf, "A multi-view approach to audio-visual speaker verification," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6194–6198.

[11] N. Zeghidour, O. Teboul, F. d. C. Quitry, and M. Tagliasacchi, "LEAF: A learnable frontend for audio classification," *arXiv preprint arXiv:2101.08596*, 2021.

[12] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNS: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[13] S. Li, Y. Yao, J. Hu, G. Liu, X. Yao, and J. Hu, "An ensemble stacked convolutional neural network model for environmental event sound recognition," *Applied Sciences*, vol. 8, no. 7, 2018.

[14] J. Huang, H. Lu, P. Lopez Meyer, H. Cordourier, and J. Del Hoyo Ontiveros, "Acoustic scene classification using deep learning-based ensemble averaging," in *2019 Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, New York University, NY, USA, October 2019, pp. 94–98.

[15] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 1021–1028.

[16] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," in *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, vol. 11006. International Society for Optics and Photonics, 2019, p. 1100612.

[17] L. N. Smith, "A disciplined approach to neural network hyper-parameters: Part 1 - learning rate, batch size, momentum, and weight decay," *arXiv preprint arXiv:1803.09820*, 2018.

[18] S. Wang, A. Mesaros, T. Heittola, and T. Virtanen, "A curated dataset of urban scenes for audio-visual scene analysis," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021.

[19] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2018.

[20] S. Wang, T. Heittola, A. Mesaros, and T. Virtanen, "Audio-visual scene classification: analysis of dcase 2021 challenge submissions," *arXiv preprint arXiv:2105.13675*, 2021.

[21] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," in *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2017, pp. 1821–1830.